Table of Contents

INTRODUCTION

ORGANIZATION OF THE REPORT

The state of Illinois is in the process of building a state level student unit record data system that will ultimately track individual students from pre-school to their employment. This P-20 (Pre-Kindergarten to Graduate School) data system is called the Illinois Longitudinal Data System (ILDS). The purpose of this system is to facilitate achieving the state goals identified in the Illinois Public Agenda:

1. Increase educational attainment to match best-performing U.S. states and world countries;
2. Ensure college affordability for students, families, and taxpayers;
3. Increase number of quality post-secondary credentials to meet demands of the economy;
4. Better integrate Illinois' educational, research, and innovation assets to meet the economic needs of the state and its regions

These four goals involve using the ILDS in numerous ways and one of the most important ways will be in conducting research based on the longitudinal experiences and outcomes of individuals who have records maintained in the system. This report, developed with a grant provided to DePaul University by IBHE, is intended to help those who are doing the longitudinal studies. It is not exhaustive in its content or discussions. It is intended to address key issues and

concerns and to provide pathways to additional material.

There are two parts to this report, the first part looks at using a longitudinal data system. It provides an introduction, context, and examples of longitudinal studies. The second part focuses the technical aspects of a longitudinal data system. It includes construction of a longitudinal data system, methodological concerns, the technology of storage and display, and the different sources of data that may be included. There is an Appendix that includes a glossary of the multitude of acronyms so prevalent in a technical discussion. The Appendix also includes references used in this discussion that provide additional material on the topics discussed.

While this discussion focuses on the longitudinal data system, the parts and their sections are also designed so that readers can go directly to the aspects most important to them. The following is a brief description of the two parts and their sections to aid in focusing on the aspects of the document that are most relevant to specific issues.

After this introduction, the first section of Part 1, (Section 1) provides a working taxonomy to categorize the different types of longitudinal studies that are conducted and their characteristics, comparing them to the primary alternative of cross-sectional research. The second section (Section 2) goes into more detail about the types of questions that may be addressed with longitudinal studies and the advantages of this approach, citing many

of the key scholars and practitioners of this research. The third section (Section 3) explores the context for longitudinal databases and their use in relevance to various funding initiatives, national standards, federal reporting, policy research, state oversight, and policy agendas. Examples of longitudinal studies at nation, region, state, and institution levels are the focus of Section 4, with a review of their use in scholarly studies and graduate work. Finally, the last section of Part 1 (Section 5) gives an overview of the different types of benchmarks and performance measures that may be calculated with longitudinal data, including those related to community college student success, transfers, student typologies, and assessment.

The second part of the report focuses on building and using the longitudinal database in longitudinal research. Section 6 discusses the practical aspects of building a longitudinal dataset, including extracting data, data structures, data manipulation issues that arise, data integrity, the use of student identifiers, defining cohorts, crosswalks and taxonomies for categorical variables, and tracking time to completion. Methodological concerns are presented next in Section 7, including design issues, advanced statistics, issues in instrument/survey construction, problems with definitions of "value added," program evaluation, sampling, and the use of multiple methods.

The technology of data storage and display is the topic of Section 8, including current expectations for dynamic web display, incorporation of new business

intelligence (BI) tools, open source and free alternatives for software, and observations about the information technology change process. Different sources of data and key variables of interest are highlighted in Section 9, including P-12 (Pre-Kindergarten to High School) schooling, social services, financial aid, employment, industry certifications, workforce training and non-credit instruction, and learning management systems. In conclusion, some planning issues are put forward for consideration by the reader in Section 10. Other issues such as the Family Educational Rights and Privacy Act (FERPA), data-sharing agreements, security, access, project planning, Information Technology support (to include the cloud), Business Intelligence, and data mining and visualization are beyond the scope of this monograph. While incidental mention is given, the reader is referred elsewhere; where possible, however, resources are provided for further exploration.

Throughout the report, the first occurrence of an acronym or abbreviation in the text is spelled out in each section. The Appendix in Section 11 contains a list of abbreviations, references, and a list of websites mentioned.

Part I: Context, Measures, and Examples

The numerous decisions that are made in designing Student Unit Records (SUR) and building systems for their collection require that we understand longitudinal research and determine the data elements that support the types of decision-making that is done. The purpose of this publication is to help with these needs. This presentation must start with broad questions, such as: What is longitudinal research? What work has already been done on longitudinal research about higher education? What has been learned? What are some of the tools that can be used to study questions with this approach?

The purpose of a longitudinal database will frequently drive the technical design of what data are to be included in the database, how the database is organized, how the database is managed, and who is involved in key decisions concerning the database. The core data in the Illinois longitudinal database are established by the Higher Education Consortium. It is envisioned that data will be integrated with other data across the educational experiences of students. As these data are shared with institutions, it is expected that the institutions will augment the core data with local, contextual data to inform institutional decision-making. As the institution creates its own data mart of student data, it is important that its staff consider the alternatives and uses of the longitudinal data. The following discussions are intended to help with the decisions that an institutions needs to

make in selecting and managing its student longitudinal databases.

SECTION 1. TYPES OF LONGITUDINAL STUDIES

Based on the approaches outlined above and this review, a working list of longitudinal study and report types may be developed. These include:

1. National, sample studies (ex. National Center for Education Statistics (NCES), the Beginning Postsecondary Students(BPS) and the Wabash National Study Surveys);
2. National, population studies using the National Student Clearinghouse for special topics (ex. national attainment rates, transfers)
3. National, federal population studies as part of oversight of financial aid (ex. Free Application for Federal Student Aid (FAFSA), TRIO Programs, scholarships, Gainful Employment)
4. Quantitative studies for research (ex. *Research in Higher Education* articles, dissertations, and the work of Pascarella and Terenzini);
5. Analysis of unit record level data collected by state agencies for policy analysis (ex. affect of a new financial aid, grant policy);
6. Institutional studies to understand persistence and completion, program evaluation, assessment,

and student achievement and institutional effectiveness for accreditation purposes;

7. Unit record studies of one or more states/systems conducted by policy organizations, associations, foundations, and others to improve student success or to monitor research grants and scholarships (ex. Achieve the Dream (ATD), Jobs for the Future, Gates);

8. State longitudinal data systems (SLDS) focused on the continuum from P-12 to the workforce;

9. Reports based on categorical characteristics about cohort progression and achieving outcomes;

10. Other purposes not listed above.

While this monograph will provide some support for all of these purposes, its primary focus is with the sixth purpose – using the SLDS in institution studies to examine a variety of questions and interests such as retention, interventions, value-added college impact, and institutional effectiveness expected as part of regional, national, and disciplinary accrediting agencies. Yet the same data may be of use for a dissertation; or a sampling frame may be necessary due to the use of a particular instrument to ensure stratification. It is helpful to think about studies done that are comparable to one's one and to look for effective practices, such as how previous researchers have addressed the vagaries of defining transfer and intent to complete.

A TAXONOMY OF LONGITUDINAL STUDIES

The term "longitudinal study" can imply several different methodological designs and reasons for research. The purpose of this section is to assist researchers in using the data that are collected as part of a specific state longitudinal data system (SLDS) initiative. However, there are many ways in which longitudinal studies may be approached. These approaches may be thought of in terms of: (1) organizational focus; (2) topic/question; (3) respondents; (4) nature of inquiry; (5) quantitative methodology; (6) use of cohorts; (7) time periods examined; (8) granularity/levels of aggregation; (9) use of the research-driven, empirical knowledge base; (10) use of questionnaires, surveys, and instrumentation; (11) mandate for development; (12) use of multiple types data within a general source; (13) use of multiple data sources across the continuum; (14) particular data structure and storage; and (15) technology for dissemination. The following gives additional detail on the types of studies developed under these different approaches and also gives examples of some of the studies that have been done.

Approach	Types	Examples
Organizational focus	Nation, region, multi-state, state, system, sector, control, institution, program	NCES BPS Longitudinal study, Wabash, SURE, Multi-State "Human Capital Development Data System, NSC Signature Reports, retention committee
Topic/Question	What is the primary topic or question of study?	Students, faculty, financial aid, expenditures, revenues, publications, research
Respondents	Entire population, sample, panel	Panels sampled from NPSAS & tracked over time by NCES, GRS subgroups
Nature of inquiry	Assessment, retention study, theory development, policy question, dissertation/thesis	Use of student engagement theory & surveys
Quantitative methodology	Multivariate, multiple methods, qualitative, descriptive, data mining	Cox regression, path analysis, factor analysis
Use of cohorts	Demographic/program breakouts	IPEDS Graduation Rate Survey, IPEDS retention ratios
Time periods examined	Cross-sectional with different groups at same time, repeated measures, multiple snapshots over time	Beginning vs. graduating students in same Spring semester, longitudinal data by semester over 10 years
Granularity/Aggregation	Organization, division, program, subgroup, or individual level view of data	Data collected for program intervention over time, broken out by cohorts; view data at program level with interventions
Research knowledge base	Based on review of cumulative literature & research, foundation agenda, policy question,	Review of Pascarella & Terenzini, Ewell, RHE, AIR Professional File & IR Resources, Tinto, others; Lumina Big Goal, ATD, JFF, NGA, CCA[1]
Use of questionnaires, surveys, & instrumentation	Vendor, research-driven, home-grown, none	Use HERI, CLA, NSSE, Holland, etc. vs. doing analysis based on existing research & theory
Mandate for development	SLDS requirement of Federal/ARRA stimulus monies, WIA, state SUR data collection, IPEDS GRS, VFA, VSA	Build to meet funding requirement, as part of grant, to meet association expectations, to meet legislative requirements

---

[1] These and other acronyms are explained in the Glossary in the Appendix.

Multiple data types within source	Ex. in post-sec: admissions/ testing, enrollment, financial aid, course-taking, awards, licensure, learning communities, finance	Bring together different sources for derived, value-added variables of interest, such as attaining milestones, Tipping Points
Linking Data sources across the continuum	Linking different data sets; movement from education to employment; Difficult to get P-12 & post-sec together without standard student identifier	P-20 to workforce, social services to post-sec, none
Data structure & storage	Complex relational model requiring query tools, functional tables or views in data warehouse, data mart with single data source	Data warehouse in Oracle with OBIEE tools
Technology of dissemination	Data warehouse/mart reports, dashboards, business intelligence tools	Visual Tableau, Zogo, Micro-Strategy

While the national call to action and policy context for accountability may be heard differently at the institution level, most of the issues that must be addressed with longitudinal studies of this nature are the same.

Bauer (2004) presents four types of longitudinal designs, citing Menard (1991). The first type, *total population designs*, involves a study of an entire population over time in different time periods. It is understood that the number of records/cases will change with death, dropout, and other reasons; but group change and trends may be examined. *Cross-sectional designs* are the second type and are the most popular, with samples/cases drawn at one or more times. Groups are included that are at different stages of maturity. For example, when an entire group of students are studied and results are interpreted based on a freshman to senior student level, this is a cross-sectional design. Examples include the College Student Experiences

Questionnaire (CSEQ) and the Higher Education Research Institute's (HERI) First Year College Survey. With this design, it is possible to study aggregate trends during one period. However, this can't be used to "study developmental patterns within a cohort and to examine causal relationships" (Bauer, 2004, p. 78). The cross-sectional design is a "'snapshot' of the influences at a single point in time" (Terenzini, 1987, p. 28). However, "Differences found between or among groups may be due to differences between or among the groups at the time they enrolled. Failure to consider such pre-college differences may lead the researcher to conclude (unwittingly and perhaps expensively) that the sources of attrition lie within the institutions control when, in fact, they do not" (p. 28). The strength of the cross-sectional study is that it can be done at a point in time. The limitation of the cross-sectional study is that assumptions must be made that the groups represent sets of similar